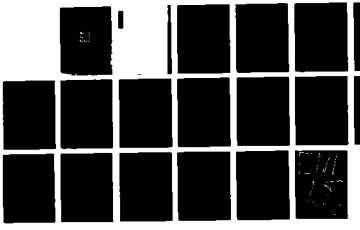
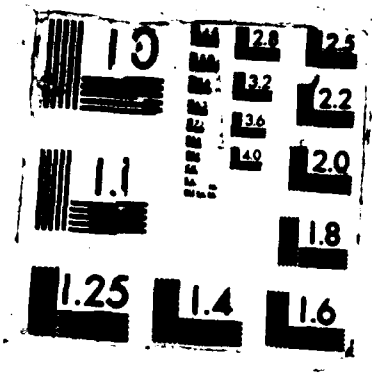
AD-A183 533 CONSISTENCY OF AKAIKE'S INFORMATION CRITERION FOR INFINITE VARIANCE AUTOREGRESSIVE PROCESSES(U) WASHINGTON UNIV SEATTLE DEPT OF STATISTICS K KNIGHT JAN 87 TR-97 N00014-84-C-0169 F/G 12/3 1/1 NL

UNCLASSIFIED




1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

Consistency of Akaike's Information Criterion for Infinite Variance Autoregressive Processes

Keith Knight

University of British Columbia and University of Washington

ABSTRACT

Suppose $\{X_n\}$ is a p-th order autoregressive process with innovations in the domain of attraction of a stable law and the true order p unknown. The estimate of p, $\hat{p}$, is chosen to minimize Akaike's Information Criterion over the integers $0, 1, \ldots, K$. It is shown that $\hat{p}$ is weakly consistent and the consistency is retained if $K \to \infty$ as $N \to \infty$ at a certain rate depending on the index of the stable law.

March 14, 1987

This research was supported by ONR Contract N00014-84-C-0169 and by NSF Grant MCS 83-01807 and was part of the author's Ph.D. dissertation completed at the University of Washington. The author would like to thank his supervisor R. Douglas Martin for his support and encouragement in preparing this paper.

0. Introduction

Consider a stationary p-th order autoregressive (AR(p)) process $\{X_n\}$:

$$X_n = \beta_1 X_{n-1} + \beta_2 X_{n-2} + \cdots + \beta_p X_{n-p} + \varepsilon_n$$

where $\{\varepsilon_n\}$ are independent, identically distributed (i.i.d.) random variables. The parameters $\beta_1, \ldots, \beta_p$ satisfy the usual stationarity constraints, namely all zeroes of the polynomial

$$z^p - \sum_{j=1}^{p} \beta_j z^{p-j}$$

have modulus less than 1.

Now assume that the true order p is unknown but bounded by some finite constant $K(N)$. Our main purpose here will be to estimate p by $\hat{p}$ where $\hat{p}$ will be obtained by minimizing a particular version of Akaike's Information Criterion (AIC) (Akaike, 1973) over the integers $\{0, 1, \ldots, K(N)\}$. Because we should be willing to examine a greater range of possible orders for our estimate as the number of observations increases, it makes sense to allow $K(N)$ to increase with N. In the finite variance case with $K(N) \equiv K$, AIC does not give a consistent estimate of p; in fact, there exists a nondegenerate limit distribution of $\hat{p}$ concentrated on the integers $p, p+1, \ldots, K$ (see Shibata, 1976).

It should be noted that AIC is a very general procedure which applies to a variety of statistical models. In general, for a given statistical model Ω_b with k-dimensional parameter vector $\underline{b}$, AIC is defined as follows:

$$\phi(\Omega_b) = -2\,\Lambda(\hat{\underline{b}}) + 2k$$

where $\Lambda(\hat{\underline{b}})$ is the maximized log-likelihood for the model Ω_b. However, in the time

series literature, AIC is usually defined in terms of a Gaussian likelihood; so for a k-th order autoregressive model, we will define AIC as follows:

$$\phi(k) = N \ln \hat{\sigma}^2(k) + 2k$$

where $\hat{\sigma}^2(k)$ is the estimate of the innovations variance obtained from the YW estimating equations. We will choose as our estimate of p the order which minimizes $\phi(k)$ for k between 0 and K, that is,

$$\hat{p} = \arg \min_{0 \le k \le K} \phi(k) .$$

In the case where two or more orders achieve the minimum, we will take the smallest of those to be our estimate.

For certain reasons, we may also want the autoregressive parameters to vary (with N) over some region of the parameter space. For example, consider the following hypothesis testing problem:

$$H_0 : X_n = \varepsilon_n$$

versus

$$H_a : X_n \text{ is a nondegenerate autoregressive process.}$$

We can consider a sequence of local alternatives $\{ H_a^{(N)} \}$ converging to H_0 in the sense that all the AR parameters converge to zero and then investigate the power of AIC as a statistical test.

The set of parameters which obey the stationarity condition is a complicated region in R^p (although the closure of this region is a compact set in R^p). However, it can be shown (Barndorff-Nielsen and Schou, 1973) that there exists a one-to-one continuous mapping

between the set of β's and the set of the first p partial autocorrelations $\{(\rho_1, \ldots, \rho_p) : \rho_j \in (-1,1) \text{ for } j = 1, \ldots, p\}$. Thus one can parametrize an AR(p) process by its p partial autocorrelations, each of which may vary freely in the interval $(-1,1)$. Moreover, one can show that for an AR(p) process, $\rho_p = \beta_p$. For autoregressive order selection, the ρ-parametrization is somewhat more natural than the β-parametrization. That is, the "distance" between two autoregressive models with different orders is more easily seen in the ρ-parametrization.

1. Infinite variance autoregressions

We will be interested in the case where the innovations $\{\varepsilon_n\}$ are in the domain of attraction of a stable law with index $\alpha \in (0,2)$. If $E(|\varepsilon_n|) < \infty$ then we will assume that $E(\varepsilon_n) = 0$.

Recall that given observations $X_1, \ldots, X_N$ and known order p, it is possible to consistently estimate the AR parameters $\beta_1, \ldots, \beta_p$. In fact for LS estimates *betahatsub* $1, \ldots, \hat{\beta}_l$ where $l \geq p$:

$$N^{1/\delta}(\hat{\beta}_k(l) - \beta_k) \xrightarrow{a.s.} 0 \quad \text{for} \quad \delta > \alpha$$

where $\beta_k = 0$ for $k > p$. For YW estimates, a slightly weaker result holds: convergence to 0 is in probability rather than almost sure.

We may also wish to consider AR models of the form

$$X_n - \mu = \beta_1(X_{n-1} - \mu) + \cdots + \beta_p(X_{n-p} - \mu) + \varepsilon_n$$

where μ is unknown and we retain the same assumptions on the β_k's and $\{\varepsilon_n\}$. It can be shown (Knight, 1987) that if we center the observed series by subtracting the sample mean $\bar{X}$ (i.e., $X'_n = X_n - \bar{X}$) and estimate $\beta_1, \ldots, \beta_p$ via the YW equations (using X'_n, $n = 1, \ldots, N$), we will still have $N^{1/\delta}(\hat{\beta}_k - \beta_k) \xrightarrow{p} 0$ for $\delta > \max(1, \alpha)$ and the convergence is almost sure for LS estimates. More generally, we can center the observed series by subtracting any location estimate $\hat{\mu}$ and estimate the β's using the centered series. Depending on the precise convergence properties of $\hat{\mu}$ we may be able to obtain the full rate of convergence for the estimates of the AR parameters (Knight, 1987).

As stated earlier, we will want to vary the autoregressive parameters with N. For this reason, we will consider a triangular array of random variables

$$
\begin{array}{l}
X_1^{(1)} \\
X_1^{(2)} \ , \ X_2^{(2)} \\
\quad \vdots \\
X_1^{(N)} \ , \ X_2^{(N)} \ , \ \ldots , \ X_N^{(N)} \\
\quad \vdots \quad \vdots
\end{array}
$$

where each row is a finite realization of an AR(p) process:

$$X_n^{(N)} = \sum_{j=1}^{p} \beta_j^{(N)} X_{n-j}^{(N)} + \varepsilon_n^{(N)} \ .$$

The corresponding triangular array of innovations, $\{ \varepsilon_n^{(N)} \}_{n \le N}$, consists of row-wise independent random variables sampled from a common distribution which is in the domain of attraction of a stable law. Given a single i.i.d. sequence $\{ \varepsilon_n \}$, we could construct each element of the triangular array as follows:

$$X_n^{(N)} = \sum_{j=0}^{\infty} c_j (\underline{\beta}^{(N)}) \varepsilon_{n-j} \ .$$

We will require that $\underline{\beta}^{(N)} = (\beta_1^{(N)} , \ldots , \beta_p^{(N)})$ are such that $(\rho_1^{(N)} , \ldots , \rho_p^{(N)})$ are contained in a closed (and hence compact) subset of $(-1, 1)^p$ for all N. Since $\beta_p^{(N)} = \rho_p^{(N)}$, we can attempt to shrink $\beta_p^{(N)}$ to zero as N goes to infinity and try to consistently estimate p at the same time. (In the testing setup mentioned earlier, this would correspond to AIC providing a consistent test under a sequence of local autoregressive

alternatives.) Intuitively, it would seem that the smaller $|\beta_p^{(N)}|$ is, the more difficult it should be to distinguish between a p-th order and a lower order AR model. From simulations, this does seem to be the case. This is the real motivation for allowing the parameters to vary with N. Consider the following example. Suppose we observe a p-th order AR process which has ρ_p very close to zero (say $\rho_p = 0.1$). To estimate the order of the process, we use a procedure which we know to be consistent. So for N large enough, we will select the true order with arbitrarily high probability. However, for moderate sized N, the probability of underestimating p may be very high. Conversely, if $|\rho_p|$ is close to 1, then even for small N there will be high probability of selecting the true order. So by allowing $\rho_p = \beta_p$ to shrink to zero with N, we may get some idea of the relative sample sizes needed to get the same probability of correct order selection for two different sets of AR parameters. If we view order selection as a hypothesis testing problem (say testing a null hypothesis of white noise versus autoregressive alternatives), shrinking β_p to zero is similar in spirit to the sequence of contiguous alternative hypotheses to a null hypothesis considered in Pitman efficiency calculations.

We should note that the partial autocorrelations do not have their usual finite variance interpretation; however, they can be unambiguously defined in terms of the regular autocorrelations which themselves can be unambiguously defined in terms of the linear process coefficients. (see Davis and Resnick, 1985) Moreover, the partial autocorrelations can be estimated consistently by recursive YW estimates just as in the finite variance case.

If we include unknown location, μ, in the model, we will assume that it does not vary with N. To have μ vary with N does not really make a lot of sense since it is, in a sense,

a nuisance parameter in this situation.

We will provide an answer to the following question: under what conditions (if any) on $K(N)$ and $(\beta_1^{(N)}, \ldots, \beta_p^{(N)})$ will AIC provide a consistent estimate $\hat{p}$ of p ? Bhansali (1983) conjectures that AIC may provide a consistent estimate of the order of an autoregressive process based on the rapid convergence of parameter estimates. However, he seems to conclude, from Monte Carlo results, that this may not be the case. If $K(N)$ is allowed to grow too fast then we may wind up severely overfitting much of the time; for example, $\hat{p}$ could equal $K(N)$ with high probability.

2. Theoretical Results

The main result of this paper is contained in Theorem 7; the first six results provide the necessary machinery for Theorem 7. We begin by stating two results dealing with r-th moments of martingales and submartingales.

Theorem 1. (Esseen and von Bahr, 1965) *Let* $S_n = \sum_{k=1}^{n} X_k$. *If* $E(X_n \mid S_{n-1}) = 0$ for $2 \le n \le N$ and $X_n \in L^r$ for $1 \le r \le 2$ then

□ $$E(|S_N|^r) \le 2 \sum_{n=1}^{N} E(|X_n|^r).$$

(Note that $\{S_n , \sigma(S_n) ; n \ge 1\}$ is a martingale.)

Theorem 2. (c.f. Chung,1974 p.346) If $\{X_n , \sigma(X_n) ; n \ge 1\}$ is an L^r-submartingale for some $r > 1$ then

□ $$E\left[\max_{1 \le n \le N} |X_n|^r \right] \le \left[\frac{r}{r-1} \right]^r E(|X_N|^r).$$

The following lemma will allow us to ignore the dependence on N of the moments of $\{X_n^{(N)}\}$ by virtue of being able to bound the moments over any sequence of admissible parameters within a compact set.

Lemma 3. Let $\{X_n(\underline{\beta})\}$ be a stationary AR(p) process with parameter $\underline{\beta}$ and innovations $\{\varepsilon_n\}$ in the domain of attraction of a stable law with index α. Let C be a compact set of the parameter space. Then for all $0 < \delta < \alpha$,

$$\sup_{\beta \in C} E\left[|X_n(\underline{\beta})|^{\delta} \right] < \infty .$$

Proof. $X_n(\underline{\beta}) = \sum_{j=0}^{\infty} c_j(\underline{\beta})\varepsilon_{n-j}$ where $c_j(\underline{\beta})$ is a continuous function of $\underline{\beta}$ for all j.

Now

$$|X_n(\underline{\beta})| \leq \sum_{j=0}^{\infty} |c_j(\underline{\beta})|\,|\varepsilon_{n-j}|$$

$$\leq \sum_{j=0}^{\infty} a_j |\varepsilon_{n-j}|$$

where $a_j = \sup_{\beta \in C} |c_j(\underline{\beta})|$. However, it can be shown that $|a_j| \leq Cj^p |x|^j$ where $|x| < 1$ and so $\sum_{j=0}^{\infty} |a_j|^{\gamma} < \infty$ for all $\gamma > 0$. Under this summability condition, it follows from Cline (1983) that the random variable

$$X = \sum_{j=0}^{\infty} a_j |\varepsilon_j|$$

is finite almost surely with

$$\lim_{x \to \infty} \frac{P[X > x]}{P[|\varepsilon_1| > x]} = \sum_{j=0}^{\infty} a_j^{\alpha} < \infty .$$

This implies that $E(X^{\delta})$ is finite for all $0 < \delta < \alpha$ and the result follows. □

The following lemma will allow us to treat moments of $\sum X_n$ the same as the moments of $\sum \varepsilon_n$ when $\alpha > 1$.

Lemma 4. Let $\{X_n\}$ be a zero mean stationary AR(p) process with innovations $\{\varepsilon_n\}$ in the domain of attraction of a stable law with index $\alpha > 1$. Then for any $1 < r < \alpha$,

(a) $E\left[\left|\sum_{n=1}^{N} X_n\right|^r\right] = O(N)$

(b) $E\left[\max_{1\le m\le N}\left|\sum_{n=1}^{m} X_n\right|^r\right] = O(N)$.

Proof.

$$\sum_{n=1}^{N} \varepsilon_n = \sum_{n=1}^{N}\left[X_n - \sum_{k=1}^{p} \beta_k X_{n-k}\right]$$

$$= \left[1 - \sum_{k=1}^{p} \beta_k\right] \sum_{n=1}^{N} X_n + R_N$$

where $|R_N| \le \left[\max_{1\le k\le p} |\beta_k|\right] p(p+1) \left[\max_{1-p\le k\le N} |X_k|\right]$. Thus

$$\sum_{n=1}^{N} X_n = C\left[\sum_{n=1}^{N} \varepsilon_n - R_N\right]$$

where $C = \left[1 - \sum_{k=1}^{p} \beta_k\right]^{-1}$. Thus by Minkowski's Inequality,

$$E\left[\left|\sum_{n=1}^{N} X_n\right|^r\right]^{1/r} \le C\, E\left[\left|\sum_{n=1}^{N} \varepsilon_n\right|^r\right]^{1/r} + C\, E[\,|R_N|^r\,]^{1/r}$$

and part (a) follows from Theorem 1 by noting that $E[\,|R_N|^r\,] = O(N)$. (It can actually be shown to be $o(N)$ by using a uniform integrability argument.)

Part (b) follows similarly from Theorem 2 by noting that

$$\max_{1\le m\le N}\left|\sum_{n=1}^{m} X_n\right| \le C \max_{1\le m\le N}\left|\sum_{n=1}^{m} \varepsilon_n\right| + C\, R_N$$

and using Minkowski's Inequality. □

The following theorem deals with uniform convergence of both LS and YW autoregressive parameter estimates in the case where location is known.

Theorem 5. Assume known location μ. Let $K(N) = O(N^{\delta})$ for $\delta < \frac{2-\alpha}{2}$ and let $\|\underline{v}\|$ denote the Euclidean norm of the vector $\underline{v}$. Then

(a) $\sqrt{N} \max_{p \le l \le K(N)} \|\underline{\hat{\beta}}(l) - \underline{\beta}^{(N)}\| \xrightarrow{p} 0$ where $\beta_k^{(N)} \equiv 0$ for $k > p$.

(b) $\sqrt{N} \max_{1 \le l \le K(N)} \|\underline{\hat{\beta}}(l) - \underline{\tilde{\beta}}(l)\| \xrightarrow{p} 0$.

Note that the vectors are not fixed length but may vary with N.

Proof. (a) The style of proof will mimic Hannan and Kanter (1977). For convenience we suppress the notation indicating the dependence of $\{X_n\}$, $\{\varepsilon_n\}$ and $\underline{\beta}$ on N. For $l \ge p$ the LS estimating equations can be reexpressed as follows:

$$\tilde{C}_l \, (\underline{\hat{\beta}}(l) - \underline{\beta}) = \underline{r}_l^*$$

where $\underline{r}_l^*(j) = \sum_{n=l+1}^{N} \varepsilon_n X_{n-j}$. Fix $\delta < \frac{2-\alpha}{2}$ and set $K = K(N) = O(N^{\delta})$. For each l, $\tilde{C}_l$ is non-negative definite and so it suffices to show that for some $\kappa < \frac{2}{\alpha}$,

(i) $\max_{p \le l \le K} N^{\frac{1}{2}-\kappa} \|\underline{r}_l^*\| \xrightarrow{p} 0$

(ii) $\min_{p \le l \le K} \min_{\|\underline{v}\|=1} N^{-\kappa} \underline{v}' \tilde{C}_l \underline{v} \xrightarrow{p} \infty$

where $K = K(N)$. If (i) and (ii) hold then clearly $\sqrt{N} \max_{p \le l \le K} \|\underline{\hat{\beta}}(l) - \underline{\beta}\| \xrightarrow{p} 0$.

To prove (i), it suffices to show that

$$E_N = E\left[\max_{p\le l\le K}\left[N^{1-2\kappa}\sum_{j=1}^{l}\left|\sum_{n=l+1}^{N}\varepsilon_n X_{n-j}\right|^2\right]^{\gamma}\right] \to 0$$

for some $\gamma < \frac{\alpha}{2}$.

Now

$$E_N \le \sum_{j=1}^{K} N^{(1-2\kappa)\gamma} E\left[\max_{1\le l\le K}\left|\sum_{n=l+1}^{N}\varepsilon_n X_{n-j}\right|^{2\gamma}\right]$$

$$\le \sum_{j=1}^{K} N^{(1-2\kappa)\gamma} E\left[\left\{\max_{1\le l\le K}\left|\sum_{n=1}^{l}\varepsilon_n X_{n-j}\right|^{\gamma} + \left|\sum_{n=1}^{N}\varepsilon_n X_{n-j}\right|^{\gamma}\right\}^2\right]$$

$$\le \sum_{j=1}^{K} N^{(1-2\kappa)\gamma}\left\{E\left[\max_{1\le l\le K}\left|\sum_{n=1}^{l}\varepsilon_n X_{n-j}\right|^{2\gamma}\right] + E\left[\left|\sum_{n=1}^{N}\varepsilon_n X_{n-j}\right|^{2\gamma}\right]\right\}$$

$$+ \sum_{j=1}^{K} N^{(1-2\kappa)\gamma}\left\{2\,E\left[\max_{1\le l\le K}\left|\sum_{n=1}^{l}\varepsilon_n X_{n-j}\right|^{2\gamma}\right]^{1/2} E\left[\left|\sum_{n=1}^{N}\varepsilon_n X_{n-j}\right|^{2\gamma}\right]^{1/2}\right\}$$

$$= \sum_{j=1}^{K} N^{(1-2\kappa)\gamma}\left[V_{N,j} + W_{N,j} + 2\,V_{N,j}^{1/2}\,W_{N,j}^{1/2}\right]$$

If $2\gamma < 1$ then by the so-called c_r-inequality

$$V_{N,j} \le E\left[\sum_{n=1}^{K}|\varepsilon_n X_{n-j}|^{2\gamma}\right] = O(K(N))$$

uniformly over j between 1 and $K(N)$.

If $2\gamma \geq 1$ then $\alpha > 1$ and so $S_{k,j} = \sum_{n=1}^{k} \varepsilon_n X_{n-j}$ is a martingale for each j. Hence $|S_{k,j}|$ is an $L^{2\gamma}$-submartingale and so by Theorems 1 and 2,

$$V_{N,j} \leq C\, E[\,|S_{K,j}|^{2\gamma}\,] \leq 2C \sum_{n=1}^{K} E[\,|\varepsilon_n X_{n-j}|^{2\gamma}\,] = O(K(N))$$

uniformly over j.

Similarly it can be shown that for all permissible values of γ, $W_{N,j} = O(N)$ uniformly over j between 1 and $K(N)$. Thus for a given sequence $K(N) = O(N^{\delta})$ by taking κ sufficiently close to $\frac{2}{\alpha}$ and γ sufficiently close to $\frac{\alpha}{2}$, we will have

$$E\left[\max_{p \leq l \leq K}\left[N^{1-2\kappa} \sum_{j=1}^{l}\left|\sum_{n=l+1}^{N} \varepsilon_n X_{n-j}\right|^2\right]^{\gamma}\right] = o(1)$$

as desired.

To prove (ii), we define $X_{n,\nu}$, $\varepsilon_{n,\nu}$ as follows:

$$X_{n,\nu} = \sum_{k=1}^{K} \nu_k X_{n-k}$$

$$\varepsilon_{n,\nu} = \sum_{k=1}^{K} \nu_k \varepsilon_{n-k}$$

with $\sum_{k=1}^{K} \nu_k^2 = 1$. It suffices to show

$$\min_{\|\underline{\nu}\|=1}\left\{N^{-\kappa} \sum_{n=K+1}^{N} X_{n,\nu}^2\right\} \xrightarrow{p} \infty .$$

Now note that $X_{n,\nu} = \sum_{j=1}^{p} \beta_j X_{n-j,\nu} + \varepsilon_{n,\nu}$. By the triangle inequality,

$$\left\{ N^{-\kappa} \sum_{n=K+1}^{N} X_{n,\nu}^2 \right\}^{1/2} \geq \left| \left\{ N^{-\kappa} \sum_{n=K+1}^{N} \left[\sum_{k=1}^{p} \beta_k X_{n-k,\nu} \right]^2 \right\}^{1/2} - \left\{ N^{-\kappa} \sum_{n=K+1}^{N} \varepsilon_{n-k,\nu}^2 \right\}^{1/2} \right|$$

Now

$$N^{-\kappa} \sum_{n=K+1}^{N} \left[\sum_{k=1}^{p} \beta_k X_{n-k,\nu} \right]^2 \leq \sum_{j=1}^{p} \beta_j^2 \sum_{k=1}^{p} N^{-\kappa} \sum_{n=K+1}^{N} X_{n-k,\nu}^2$$

$$= \sum_{j=1}^{p} \beta_j^2 \sum_{k=1}^{p} N^{-\kappa} \sum_{n=K+1}^{N} X_{n,\nu}^2 + o_p(1)$$

It remains only to show that $N^{-\kappa} \sum \varepsilon_{n,\nu}^2 \xrightarrow{p} \infty$. If this is true then

$N^{-\kappa} \sum X_{n,\nu}^2 \xrightarrow{p} \infty$ since the probability that this quantity stays bounded clearly must tend to zero.

$$N^{-\kappa} \sum_{n=K+1}^{N} \varepsilon_{n,\nu}^2 = N^{-\kappa} \sum_{n=K+1}^{N} \left\{ \sum_{k=1}^{K} \nu_k^2 \varepsilon_{n-k}^2 + 2 \sum\sum_{1 \leq j < k \leq K} \nu_j \nu_k \varepsilon_{n-j} \varepsilon_{n-k} \right\}$$

Now

$$\sum_{n=K+1}^{N} \sum_{k=1}^{K} \nu_k^2 \varepsilon_{n-k}^2 = \sum_{k=1}^{K} \sum_{n=K+1}^{N} \nu_k^2 \varepsilon_{n-k}^2$$

$$\geq \sum_{k=1}^{K} \nu_k^2 \sum_{n=K+1}^{N-K} \varepsilon_n^2$$

$$= \sum_{n=K+1}^{N-K} \varepsilon_n^2 \quad .$$

Thus

$$N^{-\kappa} \sum_{n=K+1}^{N} \sum_{k=1}^{K} v_k^2 \varepsilon_{n-k}^2 \xrightarrow{p} \infty$$

since

$$N^{-\kappa} \sum_{n=K+1}^{N-K} \varepsilon_n^2 \xrightarrow{p} \infty .$$

Thus we need only show that

$$N^{-\kappa} \sum_{n=K+1}^{N} \sum\sum_{1 \le j < k \le K} v_j v_k \varepsilon_{n-j} \varepsilon_{n-k} \xrightarrow{p} 0 .$$

Now

$$N^{-\kappa} \left| \sum_{k=2}^{K} v_k \sum_{j=1}^{k-1} v_j \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right| \le N^{-\kappa} \sum_{k=2}^{K} |v_k| \sum_{j=1}^{k-1} |v_j| \left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right| .$$

Now take $\gamma < \alpha$ and note that $j \neq k$. If $\gamma < 1$ then

$$E\left[\left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right|^{\gamma} \right] \le E\left[\sum_{n=K+1}^{N} |\varepsilon_{n-j} \varepsilon_{n-k}|^{\gamma} \right] = O(N) .$$

If $\gamma \ge 1$ then necessarily $\alpha > 1$. Thus $S_l = \sum_{n=K+1}^{l} \varepsilon_{n-j} \varepsilon_{n-k}$ is an L^{γ}-martingale and

hence

$$E\left[\left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right|^{\gamma} \right] = O(N)$$

uniformly over $j \neq k$ by Theorem 1.

Now

$$E\left[\left[N^{-\kappa}\sum_{k=2}^{K}|v_k|\sum_{j=1}^{k-1}|v_j|\left|\sum_{n=K+1}^{N}\varepsilon_{n-j}\varepsilon_{n-k}\right|\right]^{\gamma}\right] = O(N^{-\kappa\gamma}K(N)^{\gamma}) = o(1)$$

since $|v_k| \le 1$ for all k.

(b) From the definitions of $\hat{C}_l$, $\tilde{C}_l$, $\underline{\hat{r}}_l$ and $\underline{\tilde{r}}_l$, it is easy to see that

$$\text{(1a)} \qquad T_N = \max_{1\le l\le K}\max_{1\le i,j\le l}|\hat{C}_l(i,j)-\tilde{C}_l(i,j)| \le \sum_{n=1}^{K}X_n^2 + \sum_{n=N-K+1}^{N}X_n^2$$

and

$$\text{(1b)} \qquad S_N = \max_{1\le l\le K}\max_{1\le i\le l}|\hat{r}_l(i)-\tilde{r}_l(i)| \le \sum_{n=1}^{K}X_n^2 \ .$$

Thus using equations (1a) and (1b), we have

$$\text{(2a)} \qquad N^{-\kappa}T_N = o_p(N^{-1})$$

and

$$\text{(2b)} \qquad N^{-\kappa}S_N = o_p(N^{-1})$$

for $\kappa < \frac{2}{\alpha}$. Now using some elementary facts about vector and matrix norms and equations (2a) and (2b), we get

$$\text{(3a)} \qquad \max_{1\le l\le K}N^{-\kappa}\|\hat{C}_l-\tilde{C}_l\| = o_p(K(N)N^{-1}) = o_p(1)$$

and

$$\text{(3b)} \qquad \max_{1\le l\le K}N^{-\kappa}\|\underline{\hat{r}}_l-\underline{\tilde{r}}_l\| = o_p(K(N)^{1/2}N^{-1}) = o_p(1/\sqrt{N})$$

where the matrix norm is that which corresponds to the Euclidean vector norm.

Now from the definitions of $\hat{\underline{\beta}}(l)$ and $\tilde{\underline{\beta}}(l)$, we get

$$N^{-\kappa}\,\hat{C}_l\,(\,\tilde{\underline{\beta}}(l) - \hat{\underline{\beta}}(l)\,) = o_p\,(1/\sqrt{N}\,)$$

uniformly in l by equation (3b). Finally we must show that the minimum eigenvalue of $N^{-\kappa}\,\hat{C}_l$ tends in probability to infinity uniformly in l since $\|(N^{-\kappa}\,\hat{C}_l\,)^{-1}\|$ is (in the case of symmetric positive definite matrices) merely the reciprocal of this minimum eigenvalue. Note that for unit vectors $\underline{v}$

$$N^{-\kappa}\,\underline{v}'\,\hat{C}_l\,\underline{v} = N^{-\kappa}\,\underline{v}'\,\tilde{C}_l\,\underline{v} + N^{-\kappa}\,\underline{v}'(\,\hat{C}_l - \tilde{C}_l\,)\,\underline{v}$$

$$\geq N^{-\kappa}\,\underline{v}'\,\tilde{C}_l\,\underline{v} - N^{-\kappa}\,\|\hat{C}_l - \tilde{C}_l\| \xrightarrow{p} \infty$$

uniformly over l and unit vectors $\underline{v}$ by condition (ii) of the proof of part (a) of this theorem and equation (3a) above. Therefore

$$\|(N^{-\kappa}\,\hat{C}_l\,)^{-1}\| \xrightarrow{p} 0$$

as required. □

In the case where we have an unknown location parameter and we estimate it with some location estimate $\hat{\mu}$, we can obtain the following corollary.

Corollary 6.

1. If $(\hat{\mu} - \mu)^2 = O_p(N^{\gamma})$ for $\gamma \leq \min\left[\left[\frac{2}{\alpha} - \frac{5}{2} + \frac{\alpha}{2}\right], 0\right]$ uniformly over all compact subsets of the parameter space, then Theorem 5 still holds. For $\alpha > 1$, $\bar{X}$ satisfies this condition.

2. If $\alpha \leq 1$ and $\hat{\mu} \equiv \bar{X}$ and $K(N) = O(N^{\delta})$ for $\delta < \frac{1}{2}$, then conclusions (a) and (b) of Theorem 5 hold.

Proof. 1. (a) Assume without loss of generality that $\mu = 0$. We can again reexpress the LS estimating equations as follows:

$$\tilde{C}_l(\underline{\hat{\beta}} - \underline{\beta}) = \underline{r}_l^*$$

where now

$$\tilde{C}_l(i\ ,j) = \sum_{j=l+1}^{N} (X_{n-i} - \hat{\mu})(X_{n-j} - \hat{\mu})$$

and

$$\underline{r}_l^*(j) = \sum_{n=l+1}^{N} \left[\varepsilon_n + \hat{\mu}\left[1 - \sum_{k=1}^{p} \beta_k \right] \right](X_{n-j} - \hat{\mu})$$

$$= \underline{r}_l^{**}(j) + (N-l)\left[1 - \sum_{k=1}^{p} \beta_k \right]\hat{\mu}^2 \ .$$

By similar methods to those used in the proof of Theorem 5, it is easy to show that for some $\kappa < \frac{2}{\alpha}$,

$$\max_{p \le l \le K} N^{\frac{1}{2}-\kappa} \| \underline{r}_l^{**} \| \xrightarrow{P} 0 .$$

(The term involving $\sum X_{n-j}$ is killed using Lemma 4.)

In addition, using the conditions on $\hat{\mu}$,

$$N^{\frac{1}{2}-\kappa} K(N) N \hat{\mu}^2 \xrightarrow{P} 0 .$$

Finally, it follows easily that

$$\min_{p \le l \le K} \min_{\|\nu\|=1} N^{-\kappa} \underline{\nu}' \tilde{C}_l \underline{\nu} \xrightarrow{P} \infty .$$

(b) Defining T_N and S_N analogously to the proof of Theorem 5, we again get that for some $\kappa < \frac{2}{\alpha}$,

$$N^{-\kappa} T_N = o_p(N^{-1})$$

and

$$N^{-\kappa} S_N = o_p(N^{-1})$$

and the rest of the proof follows as in the proof of Theorem 5.

2. Everything follows from the fact that for any $0 < \delta < \alpha$,

$$E\left[\max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} X_n \right|^{\delta}\right] = O(N)$$

which implies that

$$\max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} X_n \right| = O_p(N^{1/\delta}) .$$

So by taking δ close to α and κ close to $\frac{2}{\alpha}$, we get

$$N^{\frac{1}{2}-\kappa} K(N) \frac{1}{N} \left[\max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} X_n \right|\right]^2 \xrightarrow{p} 0$$

and conclusions (a) and (b) follow directly from this. □

Theorem 7. If $\liminf_{N \to \infty} N \left| \beta_p^{(N)} \right|^2 > 2p$ and conclusions (a) and (b) of Theorem 5 hold for some $K(N)$ then

$$\hat{p} \xrightarrow{p} p .$$

Proof. First we note that since $\hat{p}$ is integer-valued, $\hat{p} \xrightarrow{p} p$ is equivalent to $P[\hat{p} = p] \to 1$ (as $N \to \infty$). From here on, we will refer to $K(N)$ as K and to $\beta_k^{(N)}$ as β_k thus suppressing the dependence on N.

Moreover we will assume that the observations X_n are already centered; that is, we have subtracted out the location estimate $\hat{\mu}$ (if we are assuming unknown location).

We now use the fact that

$$\hat{\sigma}^2(k) = \hat{\sigma}^2(0) \prod_{l=1}^{k} (1 - \hat{\beta}_l^2(l)) \qquad \text{for } k \geq 1$$

where $\hat{\sigma}^2(0) = \frac{1}{N} \sum_{n=1}^{N} X_n^2$ and $\hat{\beta}_k(l)$ is the YW estimate of β_k $(1 \leq k \leq l)$ in an AR(l) model. Now $P[\hat{p} < p] \leq P[\min_{0 \leq k < p} \phi(k) \leq \phi(p)]$.

Since

$$\min_{0 \leq k < p} \phi(k) \geq N \sum_{l=1}^{p-1} \ln(1 - \hat{\beta}_l^2(l)) + N \ln \hat{\sigma}^2(0),$$

we can write

$$\begin{aligned} P[\hat{p} < p] &\leq P[\ln(1 - \hat{\beta}_p^2(p)) \geq -2p/N] \\ &= P[(1 - \hat{\beta}_p^2(p)) \geq \exp(-2p/N)] \\ &\leq P[N \hat{\beta}_p^2(p) \leq 2p]. \end{aligned}$$

However,

$$N \hat{\beta}_p^2(p) = (\sqrt{N} |\beta_p| + o_p(1))^2$$

and so

$$\limsup_{N\to\infty} P\left[N\,\hat{\beta}_p^2(p) \le 2p\right] = 0$$

since $\liminf \sqrt{N}\ |\hat{\beta}_p^{(N)}| > \sqrt{2p}$. Thus $P[\hat{p} < p] \to 0$.

We also have that

$$P[\,\hat{p} > p] \le P[\,\phi(k) < \phi(k-1) \text{ for some } p < k \le K\,]$$

$$\le P\left[N \min_{p<k\le K} \ln(1-\hat{\beta}_k^2(k)) < -2\right].$$

If the conclusions of Theorem 5 hold, it follows that

$$N \max_{p<k\le K} \hat{\beta}_k^2(k) \xrightarrow{p} 0$$

and hence

$$N \min_{p<k\le K} \ln(1-\hat{\beta}_k^2(k)) \xrightarrow{p} 0.$$

Therefore, $P[\hat{p} > p] \to 0$.

Thus $P[\hat{p} \ne p] \to 0$ and so $P[\hat{p} = p] \to 1$ which implies that $\hat{p} \xrightarrow{p} p$. □

The "practical" implication of this theorem is that if N is large, with high probability $\hat{p}$ will equal p provided that $|\beta_p|$ is not too small with respect to N. Or in other words, for fixed (but large) N, the probability of selecting the correct order decreases as $|\beta_p|$ decreases. Finite sample Monte Carlo results seem to bear this out.

3. Simulation results

For illustrative purposes, a small simulation study was carried out using four symmetric stable innovations distributions with $\alpha = 0.5, 1.2, 1.9$ and 2.0 (the latter being the normal distribution). The underlying processes were AR(1) processes with the AR parameter $\beta = 0.1, 0.5$ and 0.9. The sample sizes considered were 100 and 900. For $N = 100$, the maximum order K was taken to be 10 while for $N = 900$, K was taken to be 15. 100 replications were made for each of the 24 possible arrangements of α, β and N. The results of the study are given in Tables 1 through 8.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	89	0	1
1	4	91	87
2	4	3	2
3	0	1	1
4	1	0	0
5	0	3	2
6	1	1	2
7	0	0	3
8	0	0	1
9	1	1	0
10	0	0	1

Table 1: Frequency of selected order for AR(1) process. $N = 100$ $\alpha = 0.5$

Estimated order	AR parameter		
	0.1	0.5	0.9
0	0	0	0
1	93	95	91
2	0	0	0
3	0	0	1
4	0	0	0
5	0	0	0
6	0	0	0
7	4	0	0
8	1	5	2
9	0	0	1
10–15	2	0	5

Table 2: Frequency of selected order for AR(1) process. $N = 900$ $\alpha = 0.5$

The results are much as expected. We can see that for $N = 100$ and $\beta_1 = 0.1$, AIC underestimates the true order with high probability. For $N = 900$, the probabilities of selecting the true order increases over those for $N = 100$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	70	0	0
1	15	86	86
2	7	7	4
3	3	3	3
4	1	1	3
5	0	1	1
6	0	0	0
7	2	0	2
8	2	1	1
9	0	1	0
10	0	0	0

Table 3: Frequency of selected order for AR(1) process. $N = 100$ $\alpha = 1.2$

Estimated order	AR parameter		
	0.1	0.5	0.9
0	0	0	0
1	80	87	90
2	6	3	3
3	6	0	1
4	1	4	3
5	1	2	0
6	0	0	0
7	2	2	1
8	1	0	0
9	0	0	0
10–15	3	2	2

Table 4: Frequency of selected order for AR(1) process. $N = 900$ $\alpha = 1.2$

Estimated order	AR parameter		
	0.1	0.5	0.9
0	57	0	0
1	25	76	71
2	5	8	10
3	2	5	9
4	3	6	6
5	2	1	2
6	3	1	0
7	1	1	1
8	1	0	0
9	0	0	1
10	1	2	0

Table 5: Frequency of selected order for AR(1) process. $N = 100$ $\alpha = 1.9$

Estimated order	AR parameter		
	0.1	0.5	0.9
0	5	0	0
1	78	74	75
2	7	14	9
3	2	2	7
4	2	5	2
5	1	1	2
6	1	1	0
7	3	1	3
8	0	0	0
9	0	0	1
10–15	1	2	1

Table 6: Frequency of selected order for AR(1) process. $N = 900$ $\alpha = 1.9$

Estimated order	AR parameter		
	0.1	0.5	0.9
0	63	0	0
1	25	75	75
2	4	3	12
3	1	6	2
4	0	7	5
5	2	2	2
6	2	4	3
7	1	0	0
8	1	1	1
9	0	2	0
10	1	0	0

Table 7: Frequency of selected order for AR(1) process. $N = 100$ Normal distribution

Estimated order	AR parameter		
	0.1	0.5	0.9
0	0	0	0
1	83	79	80
2	3	3	11
3	4	4	3
4	4	6	0
5	2	3	3
6	0	0	0
7	0	4	0
8	4	1	1
9	0	0	2
10–15	0	0	0

Table 8: Frequency of selected order for AR(1) process. $N = 900$ Normal distribution

4. Comments

Bhansali and Downham (1977) propose a generalization of AIC. They propose to minimize $\phi'(k) = N \ln \hat{\sigma}^2(k) + \gamma k$ where $\gamma \in (0,4)$. It is easy to see from the proof of the above result that their criterion will also lead to consistent estimates of p under similar conditions on $K(N)$ and $\beta_p^{(N)}$. In fact, if $\gamma = \gamma(N) > 0$ satisfies $\gamma(N)/N \to 0$, then the criterion corresponding to $\phi''(k) = N \ln \hat{\sigma}^2(k) + \gamma(N)\, k$ will consistently estimate p. Specifically, with known location, the estimate will be consistent provided

$$\liminf_{N \to \infty} \frac{N}{\gamma(N)} \left| \beta_p^{(N)} \right|^2 > p$$

with $\gamma(N)$ bounded away from zero and with the same conditions on $K(N)$. With an appropriate choice of $\gamma(N)$, this criterion will also be consistent in the finite variance case. However if $\gamma(N)$ grows too quickly with N then the criterion may seriously underestimate the true order p in small samples in both the finite and infinite variance cases. In an application such as autoregressive spectral density estimation (assuming now finite variance), underestimation is more serious than overestimation since, if the order is underestimated, the resulting spectral density estimate may be lacking important features which may indeed exist.

Davis, R. and Resnick, S. (1985) Limit theory for moving averages of random variables with regularly varying tail probabilities. *Ann. Prob.* **13** 179–95.

Esseen, C. and von Bahr, B. (1965) Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.* **36** 299–303.

Hannan, E. and Kanter, M. (1977) Autoregressive processes with infinite variance. *J. App. Prob.* **14** 411–15.

Knight, K. (1987) Rate of convergence of centred estimates of autoregressive parameters for infinite variance autoregressions. *J. Time Ser. Anal.* **8** 51–60.

Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–26.

Current address: Dept. of Statistics, University of British Columbia, 2021 West Mall, Vancouver, B.C., V6T 1W5

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, in: *Second International Symposium on Information Theory* (eds. B. Petrov and F. Csaki) Akedemiai Kiade, Budapest 267–81.

Barndorff-Nielsen, O. and Schou, G. (1973) On the parametrization of autoregressive models by partial autocorrelations. *J. Multivar. Anal.* 3 408–19.

Bhansali, R. (1983) Order determination for processes with infinite variance, in: *Robust and Nonlinear Time Series Analysis*. (eds. J. Franke, W. Hardle, R.D. Martin), Springer-Verlag, New York, 17–25.

Bhansali, R. and Downham, D. (1977) Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* 64 547–51.

Chung, K.L. (1974) *A course in probability theory.* Academic Press, New York.

Cline, D. (1983) Infinite series of random variables with regularly varying tails. Technical Report No. 83-24, Institute of Applied Mathematics and Statistics, University of British Columbia.

END

9-87

DTIC